POLITECHNIKA POZNAŃSKA

WYDZIAŁ INFORMATYKI I TELEKOMUNIKACJI
Instytut Informatyki
ul. Piotrowo 2, 60-965 Poznań, tel. +48 61 665 2997, fax +48 61 877 1525
e-mail: office_cs@put.poznan.pl, www.cs.put.poznan.pl

Instytut Informatyki

Poznań, 18.02.2021 r.

Dr hab. inż. Aleksandra Świercz
Instytut Informatyki
Politechnika Poznańska

RECENZJA ROZPRAWY DOKTORSKIEJ DLA RADY NAUKOWEJ DYSCYPLINY INFORMATYKA TECHNICZNA I TELEKOMUNIKACJA POLITECHNIKI WARSZAWSKIEJ

Tytuł rozprawy: Szacowanie liczby powtórzeń fragmentu DNA

Autor rozprawy: mgr inż. Wiktor Kuśmirek

Promotor: dr hab. inż. Robert Nowak, prof PW

Tematyka badawcza

Recenzowana rozprawa doktorska przedstawia wyniki badań dotyczące analizy sekwencji genetycznych, w szczególności sekwencji które zawierają powtarzające się fragmenty. Prace nad odczytaniem sekwencji całego genomu dla różnych organizmów trwają od lat. Jednym z najbardziej spektakularnych osiągnięć było odczytanie sekwencji genomu ludzkiego na początku dwudziestego pierwszego wieku. Proces ten zajął wówczas około 15 lat i pochłonął 3 mld dolarów. Dzięki rozwojowi technologii, znacznie przyspieszony został proces sekwencjonowania, czyli odczytywania sekwencji DNA, a także drastycznie zmniejszone zostały koszty, umożliwiając na sekwencjonowanie genomu człowieka poniżej 1000 dolarów i w czasie nie dłuższym niż kilka dni. Obecnie na rynku znajdują się dwie wiodące technologie: sekwencjonowanie nowej generacji (NGS), które pozwala uzyskać kilkusetnukleotydowe sekwencje o bardzo dobrej jakości, oraz sekwencjonowanie trzeciej generacji (TGS), dzięki której można uzyskać długie sekwencje (do kilkudziesięciu, a nawet kilkuset, tysięcy nukleotydów) o trochę gorszej jakości. Największym wyzwaniem odczytywania sekwencji genomów są powtarzające się fragmenty sekwencji, lub sekwencje o bardzo niskim stopniu skomplikowania, co znacznie utrudnia proces asemblacji, czyli składania sekwencji genomu z krótkich fragmentów pochodzących z

sekwencjonowania. Warto zwrócić uwagę, że pomimo iż sekwencja genomu ludzkiego znana jest od kilkunastu lat, to nadal zawiera nieodczytane fragmenty, które są efektem powtórzeń.

Kolejną trudnością jest fakt, iż genomy organizmów tego samego gatunku różnią się między sobą. Zmiany mogą być krótkie: jedno- lub kilku- nukleotydowe (SNP oraz INDEL) oraz mogą obejmować obszar wielu milionów nukleotydów. W obrębie największych zmian, tak zwanych wariantów strukturalnych (SV), możemy rozróżnić delecje, duplikacje, inwersje, insercje oraz translokacje. Najczęstsze różnice są najkrótsze i najłatwiejsze do wykrycia. Dla przykładu w organizmie ludzkim występuje około 3 mln SNP, czyli średnio 1 na 1000 nukleotydów jest inny niż w genomie referencyjnym. Znacznie rzadsze zmiany SV wykrywane są dzięki nowym technologiom sekwencjonowania. Wyróżniamy dwa główne podejścia: mapowanie do sekwencji genomu referencyjnego, oraz asemblacja *de novo* bez wykorzystania genomu referencyjnego. W pierwszym podejściu sprawdzane są sparowane odczyty (PEM), odczyty nie mapujące się w całości do genomu (SR), a także badana jest głębokość pokrycia (RD). To dzięki badaniu głębokości pokrycia, a także asemblacji *de novo* można wykryć zróżnicowanie liczby kopii fragmentów DNA, tzw CNV, które jest szczególnym przypadkiem SV obejmującym delecje oraz duplikacje. Badania w tym zakresie pokazały, że niektóre fragmenty genomu są bardziej podatne na CNV, oraz że mogą wpływać na ekspresję genów. Tematyka recenzowanej pracy doktorskiej wpisuje się w tematykę odczytywania genomu *de novo* oraz wykrywania powtarzających się fragmentów badanych sekwencji.

Zakres pracy i wkład autora

W pracy doktorskiej mgr W. Kuśmirek zajął się szacowaniem liczby kopii fragmentów sekwencji w genomie. Autor przetestował narzędzia do szacowania CNV na podstawie danych sekwencjonowania cało-eksomowego (WES). Sekwencjonowanie WES polega na wstępnym wyselekcjonowaniu fragmentów DNA obejmujących eksony, powieleniu tychże fragmentów DNA, a następnie ich sekwencjonowaniu. Pozwala to na pokrycie genomu tylko w wybranych obszarach, które tworzą niejako *okna*, w genomie. Zaproponowane narzędzie SeQuiLa-cov pozwala na znacznie szybsze obliczanie głębokości pokrycia w zdefiniowanych oknach niż istniejące narzędzia. Autor rozprawy przetestował różny dobór próbek referencyjnych, które będą wykorzystane do modelowania tła przy obliczaniu głębokości pokrycia. Opracował i zaimplementował algorytm do wyboru próbek referencyjnych, który można wykorzystać w aplikacjach nie posiadających tego etapu.

Drugi nurt badawczy doktoranta dotyczył rekonstrukcji *de novo* sekwencji DNA, wykorzystując szacowanie liczby kopii fragmentów DNA na etapie tworzenia grafów A-Bruijna. Zazwyczaj narzędzia do asemblacji *de novo* nie odtwarzają fragmentów repetytywnych, ze względu na rozgałęzienia w grafie, które są przez nie tworzone, gdyż w łatwy sposób można połączyć błędne sekwencje. Opracowane i zaimplementowane narzędzie dnaasm pozwala na uzyskanie kontigów również w przypadku powtarzających się sekwencji. Ponadto, doktorant zaimplementował narzędzie do łączenia kontigów uzyskanych przy użyciu asemblera dnaasm z długimi odczytami pochodzącymi z sekwencjonowania trzeciej generacji. Pozwoliło to uzyskać dłuższe kontigi o wciąż dobrej jakości. Oba te narzędzia zostały wykorzystane w praktyce podczas sekwencjonowania *de novo* genomu tasiemca.

Ocena strony merytorycznej

Rozprawa doktorska jest oparta na cyklu pięciu publikacji, które ukazały się w wysoko punktowanych czasopismach z listy JCR. Autoreferat z przedstawioną tematyką pracy doktorskiej, opisem publikacji i osiągnięciami doktoranta zawarty został na 55 stronach. Bibliografia składa się z 103 pozycji literaturowych.

Rozdział 1 wprowadza w tematykę zagadnienia, przedstawia aktualny stan wiedzy oraz prezentuje cel rozprawy i postawione hipotezy badawcze. Opisane zostały różne sposoby sekwencjonowania oraz różnice między sekwencjonowaniem całego genomu (WGS) oraz całego eksomu (WES). Przydałoby się umieścić trochę nowsze cytowania dotyczące stopy błędów w sekwencjonowaniu trzeciej generacji niż publikacja [24] z 2015 roku, gdyż technologia ONT aktualnie ma znacznie niższy odsetek błędów niż 40%. Podobnie dla PacBio poprawiona została jakość dzięki odczytom HiFi (High Fidelity). Długie odczyty o lepszej jakości pozwalają oczywiście na uzyskanie dłuższych kontigów o bardzo wysokiej jakości. W podrozdziale 1.2.1 przedstawione zostały dwa sposoby na obliczanie głębokości pokrycia na podstawie danych WES, wraz z dostępnymi narzędziami. Następnie omówione zostały sposoby na filtrowanie regionów o nienaturalnym pokryciu, które będą zaburzały dalsze analizy. W kolejnym etapie, normalizacji głębokości pokrycia, wykorzystywane są między innymi wybrane próbki jako tło modelujące. Autor rozprawy zauważył, że nie wszystkie programy wykorzystują pule próbek referencyjnych i jako jeden z celów pracy postawił sobie przetestowanie różnego sposobu doboru próbek referencyjnych oraz implementację algorytmu doboru próbek dla aplikacji nie posiadających tego etapu.

Podrozdział 1.2.2 przedstawia drugą część zagadnień poruszanych w pracy doktorskiej obejmujących szacowanie głębokości pokrycia w trakcie asemblacji *de novo*. Przedstawione różne podejścia do wykrywania wariantów strukturalnych (SV) zostały przez autora niesłusznie sklasyfikowane jako podejścia do wykrywania CNV. CNV jest podgrupą SV, a do wykrycia CNV najlepiej posłużyć może badanie głębokości pokrycia oraz asemblacja *de novo*. Pozostałe metody potrafią wykryć jedynie delecję, natomiast nie potrafią wykryć duplikacji. Niewłaściwy zapis pojawił się także na Rysunku 1 – zamiast CNV powinno pojawić się SV. W dalszej części rozdziału doktorant opisał podejścia do asemblacji *de novo* oparte na grafach OLC oraz grafach de Bruijna. W grafie można także zapisać informację z oszacowaniem ile razy należy użyć każdej krawędzi, dzięki czemu będzie można rozwiązać także problem powtarzających się fragmentów. W kolejnej części podrozdziału autor przedstawia różne sposoby na łączenie danych z sekwencjonowania długich i krótkich odczytów. Na koniec doktorant zaprezentował swój wkład w tym obszarze, czyli nowy algorytm asemblacji de novo z wykorzystaniem informacji szacowanej liczy kopii w celu odtworzenia fragmentu powtarzającego się oraz algorytm do asemblacji de novo z połączonymi sekwencjami krótkimi i długimi.

Cele pracy oraz hipotezy badawcze zostały sformułowane zarówno na początku pracy w rozdziale 1.1 oraz na koniec rozdziałów 1.2.1 i 1.2.2. W mojej opinii skrótowe przedstawienie zaraz na początku pracy może być dla czytelnika niezrozumiałe, gdyż używane są skróty wcześniej nie wyjaśnione. Bardziej właściwym byłoby umieszczenie rozdziału 1.1 po wprowadzeniu.

Tematem rozdziału 2 jest szacowanie liczby kopii na podstawie sekwencjonowania eksomu. W pierwszej części omówiona jest publikacja P1, w której testowano algorytm doboru próbek referencyjnych do 3 różnych aplikacji, które nie posiadały takiego kroku. Autor zwięźle i klarownie przedstawił najważniejsze wnioski wypływające z badań dotyczących różnych sposobów wyboru próbek referencyjnych. W drugiej części rozdziału omówiona została publikacja P4. Doktorant zajmował się przeprowadzeniem testów wydajnościowych aplikacji SeQuiLa-cov, wykazując że wykonanie obliczeń na wielu rdzeniach znacznie przyspiesza obliczanie głębokości pokrycia w porównaniu do innych aplikacji.

W rozdziale 3 omówione zostały trzy publikacje doktoranta P2, P3, P5 – związane z asemblacją *de novo* zarówno krótkich, jak i długich odczytów. Wnioski z każdej publikacji są przedstawione jasno i klarownie. W publikacjach P2 i P3 doktorant zaproponował nowatorskie algorytmy do asemblacji *de novo*, które uwzględniają informację o głębokości pokrycia i są w stanie odtwarzać sekwencje repetytywne. Pozostałe testowane asemblery (ABySS, Velvet oraz SPAdes) w znaczącej mierze nie były w stanie wykryć powtórzeń tandemowych (Tabele 4-6 w P2). Mam zastrzeżenie odnośnie szacowania zużycia pamięci i czasu działania. Autor testował swoją aplikację na małych genomach bakteryjnych, oraz danych symulowanych, gdzie rozmiar genomu nie przekraczał 1 miliona bp. Aplikacje (dnaasm i dnaasm-link) mogłyby nie zadziałać na danych pochodzących z genomu człowieka o rozmiarze 3 miliardów bp. Szacowany czas 8h nie jest zbyt wymagający, aby móc przetestować te aplikacje i sprawdzić ich działanie w praktyce na dużych genomach. W drugiej części rozdziału omówione została ostatnia z publikacji (brakuje odwołania na stronie 28 do P5) dotycząca projektu sekwencjonowania tasiemca szczurzego. Omówione zostały kolejne etapy przygotowania odczytów krótkich oraz długich. Niezrozumiałym jest dla mnie dlaczego odfiltrowane zostały odczyty poprawnie sparowane? Czy nie powinny być raczej odfiltrowane odczyty niepoprawnie sparowane? W wyniku przeprowadzonych eksperymentów obliczeniowych udało się poprawić spójność sekwencji genomu tasiemca. W ostatniej części rozdziału przedstawiony został sposób asemblacji *de novo* genomu mitochondrialnego tasiemca z wykorzystaniem krótkich odczytów. Pojawia się pytanie, skoro genom mitochondrialny jest taki krótki (13 kbp), to czy nie można było wykorzystać odczytów Nanopore? Jeden odczyt powinien pokryć cały genom mtDNA.

W ostatnim, czwartym rozdziale doktorant omówił dalsze plany badawcze, w tym automatyzację procesu doboru próbek referencyjnych, różny dobór podzbioru regionów sekwencjonowania oraz wykorzystanie metody mapowania optycznego.

Mimo wymienionych uwag oraz pytań bardzo pozytywnie oceniam recenzowaną pracę doktorską. Tematyka, choć omówiona została jako dwie odrębne części, połączona jest wspólnym zagadnieniem wykrywania różnej liczby kopii fragmentów DNA, z jednej strony jako obliczanie głębokości pokrycia w badaniu wariantów strukturalnych CNV, z drugiej jako szacowanie liczby powtarzających się fragmentów w asemblacji *de novo*. Doktorant ma rozeznanie w tematyce sposobów wykrywania sekwencji repetytywnych. Wykazał się zarówno wiedzą z zakresu informatyki, biegłością w implementacji algorytmów, skalowalności obliczeń, czy też

teorii grafów, a także wiedzą w zakresie biologii, znajomością najnowszych technologii sekwencjonowania, charakterystyką złożoności sekwencji DNA.

Ocena od strony redakcyjnej

Na recenzowaną pracę doktorską składa się zarówno cykl publikacji oraz 20-stronicowe streszczenie. Odnośnie publikacji – nie mam uwag. Streszczenie jest napisane starannie. Doktorant ma czasami tendencję do budowania bardzo długich zdań złożonych, które nie pasują do siebie i powinny być rozbite na kilka krótszych zdań. Podział pracy na rozdziały jest prawidłowo dobrany, dzięki czemu czytelnik w łatwy sposób orientuje się, co można znaleźć w której publikacji. Mgr Kuśmirek używa w pracy słów, które w dziwny sposób zostały przetłumaczone z języka angielskiego, lub nie zostały w ogóle przetłumaczone:

- Używany w pracy często 'assembling *de novo*' z angielskiego *de novo assembly* tłumaczy się na j. polski: 'asemblacja *de novo*'
- Str 13: Whole Exome Sequencing (WES) można tłumaczyć jako 'pełnoeksomowe' sekwencjonowanie, choć częściej spotykane tłumaczenie to całoeksomowe. W pracy pojawia się dodatkowo słowo 'pełnoeksonowe', co pewnie jest literówką
- Str 32: 'Optical mapping' nie zostało przetłumaczone – powinno być: mapy optyczne lub mapowanie optyczne

Autor nie ustrzegł się także przed nielicznymi błędami gramatycznymi, składniowymi i interpunkcyjnymi:

- Str. 10: „Zaprojektować i zbadać nową, wydajną implementację algorytmu" powinno być „Zaprojektować i **przetestować** nową, wydajną implementację algorytmu"
- Str. 10: „do łączenia wyników assemblingu *de novo* krótkich odczytów przez długie odczyty" powinno być „do łączenia wyników **asemblacji** *de novo* krótkich odczytów **za pomocą** długich odczytów"
- Str 11. „Każda z wymienionych hipotez została udowodniona, rozwiązania opublikowałem" zdanie zbyt długie, powinno być rozbite : „Każda z wymienionych hipotez została udowodniona. Rozwiązania opublikowałem…"
- Str 11. „Przykładowo, niektóre funkcje genów mogą być modulowane przez zmianę liczby powtórzeń DNA, **proces** ten umożliwia…" Zdanie długie, wymaga podzielenia na 2 zdania. Ponadto nie wiadomo o jaki proces chodzi.
- Str 13: „odpowiadają częścią kodującym" powinno być „odpowiadają częściom kodującym"
- Str 14. „Istotną rolę w poznaniu takich sekwencji mają algorytmy." Nie wiadomo o jakie algorytmy chodzi
- Str 14-15 „Narzędzia różnią się czasem działania, oprócz różnych algorytmów są implementowane przy pomocy różnych języków programowania, ponadto nie wszystkie aplikacje posiadają implementacje zrównoleglenia obliczeń." Zdanie, które spokojnie może zostać podzielone na 3 odrębne. Dodatkowo środkowa część zdania wymaga zmiany, aby była zrozumiała.

- Str. 16: „nową, rozproszoną implementacja procesu" powinno być „nową rozproszoną implementację procesu"
- Str. 26: „zwiększa prawdopodobieństwo odtworzenie sekwencji" powinno być „zwiększa prawdopodobieństwo odtworzenia sekwencji"
- Str. 26: „Następnie jest budowany graf połączeń w którym wierzchołkami są kontigi a krawędziami" powinno być „„Następnie jest budowany graf połączeń, w którym wierzchołkami są kontigi a krawędziami"

Szereg błędów edycyjnych pojawiło się także w cytowaniach literaturowych, wynikających najprawdopodobniej z formatowania Latex'a. są to błędy typu: duże/małe litery.

- W [3],[5] i [9] DNA pojawia się w tytule małymi literami: 'dna'
- W [95] CNV w tytule pojawia się małymi literami: 'cnv'
- Czasopisma typu PLOS pisze się pierwszy człon dużymi literami, drugi zaczyna z dużej litery: PLOS Genetics [1], PLOS ONE [13,36], PLOS Computational Biology [42, 58]
- Czasopismo Genome Research pisze się z dużych liter [32,43,48,49,55,62,69]
- Podobnie Nucleic Acid Research [22,28,35,53]
- Czasopismo BMC Bioinformatics – drugi człon z dużej litery [47,72,86,93,100]
- Również niepoprawna pisownia czasopism: Genome Biology, Genome Research, Nature Genetics, GigaScience, Journal of Molecular, Biology, Molecular Cell
- Tytuły artykułów powinny zaczynać się z dużej litery, a kolejne wyrazy z małych liter [63,67,70,78,82,85]

Powyższe usterki nie mają znacznego wpływu na jakość i czytelność pracy i nie umniejszają jej wartości. Nie zmieniają również mojej pozytywnej oceny recenzowanego doktoratu.

Wnioski końcowe

W moim przekonaniu autor recenzowanej pracy doktorskiej wykonał, szereg skomplikowanych badań. Zaproponował, zaimplementował i przetestował algorytmy, które w znaczny sposób mogą przyspieszyć obliczenia, a także poprawić jakość uzyskiwanych sekwencji wynikowych zawierających powtórzenia. Wykazał się wiedzą z zakresu matematyki, informatyki, biologii oraz bioinformatyki, którą potrafił zastosować w praktyce. Recenzowana praca zawiera oryginalne rozwiązania problemu powtórzeń w sekwencjach DNA. Wyniki badań zostały opublikowane w czterech wysoko punktowanych czasopismach z dziedziny (*BMC Bioinformatics (2x), GigaScience, Scientific Data, BioMed Research International*) dając sumaryczny współczynnik IF = 19,581. W trzech z pięciu publikacji był pierwszym autorem z przeważającym wkładem pracy. Ponadto opublikował 10 artykułów w materiałach konferencyjnych (*International Society for Optics and Photonics*), występował na konferencjach jako referent lub z prezentacją plakatową. Na podkreślenie zasługuje również fakt, iż doktorant uzyskał w 2020 roku grant Preludium przyznany przez Narodowe Centrum Nauki,

oraz grant CYBERIADA-1 finansowany przez Politechnikę Warszawską. Uczestniczył również w 4 grantach badawczych oraz różnych kursach o głębokiej analizie danych biomedycznych.

Stwierdzam, że praca pana mgr inż. Wiktora Kuśmirka pt. „Szacowanie liczby powtórzeń fragmentu DNA" spełnia wymagania stawiane rozprawom doktorskim określone w Ustawie o stopniach naukowych i tytule naukowym oraz stanowi oryginalne rozwiązanie problemu naukowego. Wnoszę o dopuszczenie mgr inż. Wiktora Kuśmirka do dalszych etapów przewodu doktorskiego.
Mając na uwadze bogaty dorobek publikacyjny zgromadzony w ciągu 4 lat pracy nad doktoratem oraz fakt iż zostały opublikowane w wiodących czasopismach naukowych, składam wniosek o wyróżnienie tej pracy doktorskiej.

Dr hab. inż. Aleksandra Świercz